Balanced Augmented Lagrangian Method for Convex Programming

Bingsheng He[1] **Xiaoming Yuan**[2]

August 17, 2021

Abstract. We consider the convex minimization model with both linear equality and inequality constraints, and reshape the classic augmented Lagrangian method (ALM) by balancing its subproblems. As a result, one of its subproblems decouples the objective function and the coefficient matrix without any extra condition, and the other subproblem becomes a positive definite system of linear equations or a positive definite linear complementary problem. The balanced ALM advances the classic ALM by enlarging its applicable range, balancing its subproblems, and improving its implementation. We also extend our discussion to two-block and multiple-block separable convex programming models, and accordingly design various splitting versions of the balanced ALM for these separable models. Convergence analysis for the balanced ALM and its splitting versions is conducted in the context of variational inequalities through the lens of the classic proximal point algorithm.

Keywords: Convex programming, augmented Lagrangian method, proximal point algorithm, proximity

1 Introduction

The classic augmented Lagrangian method (ALM) was proposed in [25,30], and since then it has been playing fundamental roles in algorithmic design for various convex programming problems. For instance, it is the root of the alternating direction method of multipliers (ADMM) proposed in [13], which is nowadays a benchmark algorithm used widely in many areas. We refer to, e.g. [3,6,12,14,32], for insightful discussions on the ALM and its wide applications in different areas such as PDEs, optimization, optimal control, image processing, and scientific computing. In particular, it was shown in [32] that the ALM is an application of the classic proximal point algorithm (PPA) which was originally proposed in [27].

Let us start with the following canonical convex minimization model with linear equality constraints:

$$\min\{\theta(x) \mid Ax = b,\, x \in \mathcal{X}\}, \tag{1.1}$$

where $\theta : \Re^n \to \Re$ is a closed proper convex but not necessarily smooth function; $\mathcal{X} \subseteq \Re^n$ is a closed convex set; $A \in \Re^{m\times n}$; and $b \in \Re^m$. The iterative scheme of ALM for (1.1) reads as

$$\text{(ALM)} \quad \left\{ \begin{array}{ll} x^{k+1} \in \arg\min\bigl\{\theta(x) - (\lambda^k)^T(Ax-b) + \frac{r}{2}\|Ax-b\|^2 \bigm| x \in \mathcal{X}\bigr\}, & \text{(1.2a)} \\ \lambda^{k+1} = \lambda^k - r(Ax^{k+1} - b), & \text{(1.2b)} \end{array} \right.$$

in which $r > 0$ is the penalty parameter and $\lambda \in \Re^m$ is the Lagrange multiplier. Hereafter, x and λ are referred to the primal and dual variables, respectively. In general, the subproblem (1.2a) needs to be solved iteratively and thus outer-inner nested iterations are rendered to implement the ALM (1.2). Therefore, how to solve the x-subproblem (1.2a) determines the difficulty of implementing the ALM (1.2). An obvious obstacle is that the objective function $\theta(x)$, the coefficient matrix A, and the set $\mathcal{X}$ are all aggregated to be considered simultaneously in the

[1] Department of Mathematics, Nanjing University, Nanjing, China. This author was supported by the NSFC Grant 11871029. Email: hebma@nju.edu.cn

[2] Department of Mathematics, The University of Hong Kong, Hong Kong, China. Email: xmyuan@hku.hk

x-subproblem (1.2a). Thus, the x-subproblem (1.2a) dominates the computation while the λ-subproblem (1.2b) is trivial. In this sense, these two subproblems in the classic ALM (1.2) are unbalanced.

In this paper, we suggest to decouple the objective function $\theta(x)$ and the coefficient matrix A in the subproblem (1.2a) so as to alleviate this subproblem substantially, and then shift the consideration of the matrix A to the subproblem (1.2b). The classic ALM (1.2) is thus reshaped, and the resulting subproblems are balanced in the sense that the difficulty of the x-subproblem only depends on $\theta(x)$ and $\mathcal{X}$, and that of the λ-subproblem becomes to depend on A. This balancing idea has an immediate advantage when the function $\theta(x)$ has the favorable property that its proximity operator can be represented by a closed-form. That is, the proximity operator of the objective function $\theta(x)$, which is defined by

$$\text{Prox}_{\theta}^{r}(x) := \arg\min\big\{\theta(y) + \frac{r}{2}\|y - x\|^2 \mid y \in \Re^n\big\},\ \forall x \in \Re^n,\ \forall r > 0, \tag{1.3}$$

has a closed-form representation. This scenario arises in many applications, especially in contemporary data science domains. We refer to, e.g., [4, 8, 31], for some applications whose corresponding function $\theta(x)$ usually prompts sparsity- or low-rank properties of a desired solution and hence can be specified as the l_1-norm function (or the nuclear-norm function for the case with matrix variables). Our idea of decoupling $\theta(x)$ and A can be further explained by the following motivation. Ignoring some constant terms, we know that the subproblem (1.2a) can be rewritten as

$$x^{k+1} \in \arg\min\big\{\theta(x) + \frac{r}{2}\|Ax - b - \frac{1}{r}\lambda^k\|^2 \mid x \in \mathcal{X}\big\}.$$

For the classic ALM (1.2), even when $\text{Prox}_{\theta}^{r}(x)$ can be represented by a closed-form and $\mathcal{X} = \Re^n$, in general the subproblem (1.2a) may still be difficult when the matrix A is not identity. If $\theta(x)$ and A are decoupled and the primeval x-subproblem (1.2a) is replaced by an easier one in form of

$$x^{k+1} = \arg\min\big\{\theta(x) + \frac{r}{2}\|x - q^k\|^2 \mid x \in \mathcal{X}\big\}, \tag{1.4}$$

in which $q^k \in \Re^n$ is a certain constant vector, then the solution of (1.4) can also be given by the closed-form representation of $\text{Prox}_{\theta}^{r}(x)$ when $\mathcal{X} = \Re^n$.

Some existing algorithms in the literature can be applied to (1.1), and $\theta(x)$ and A can be decoupled in their implementations. For example, as analyzed in [20], we can consider regularizing the objective function of (1.2a) with a proximal term. The resulting proximal version of the ALM (PALM for short) can be written as

$$\text{(PALM)}\left\{\begin{array}{ll} x^{k+1} \in \arg\min\big\{\theta(x) - (\lambda^k)^T(Ax - b) + \frac{r}{2}\|Ax - b\|^2 + \frac{1}{2}\|x - x^k\|_G^2 \mid x \in \mathcal{X}\big\}, & \text{(1.5a)} \\ \lambda^{k+1} = \lambda^k - r(Ax^{k+1} - b), & \text{(1.5b)} \end{array}\right.$$

with $G \in \Re^{n\times n}$ and the notation $\|x\|_G^2 := x^T G x$. If we choose $G = \sigma I_n - rA^T A$ in (1.5a) with $\sigma > 0$, then the generic PALM (1.5) is specified as

$$\text{(LALM)}\left\{\begin{array}{ll} x^{k+1} \in \arg\min\big\{\theta(x) + \frac{\sigma}{2}\big\|x - (x^k + \frac{1}{\sigma}A^T(\lambda^k - r(Ax^k - b)))\big\|^2 \mid x \in \mathcal{X}\big\}, & \text{(1.6a)} \\ \lambda^{k+1} = \lambda^k - r(Ax^{k+1} - b), & \text{(1.6b)} \end{array}\right.$$

in which the subproblem (1.6a) is in form of (1.4) and thus it is reduced to $\text{Prox}_{\theta}^{r}(x)$ when $\mathcal{X} = \Re^n$. The scheme (1.6) is called the linearized ALM (LALM for short) because the quadratic term $\frac{r}{2}\|Ax - b\|^2$ in (1.5a) is "linearized" by $G = \sigma I_n - rA^T A$. From an analytical point of view, it is easy to see that if σ is large enough such that $\sigma > r\|A^T A\|$, then the matrix $G = \sigma I_n - rA^T A$

is positive definite, and essentially convergence of the LALM (1.6) can be conducted by following existing works such as [11, 19, 34, 35]. This means the classic ALM (1.2) can be revised as the LALM (1.6) to decouple $\theta(x)$ and A. But the subproblem (1.6a) is correlated implicitly with A because of the condition $\sigma > r\|A^TA\|$. On the other hand, note that the approximation of (1.6a) to the primeval x-subproblem (1.2a) is less accurate when σ is larger, because of the higher weight of the additional quadratic term $\frac{1}{2}\|x-x^k\|_G^2$ with $G = \sigma I_n - rA^TA$. When $\|A^TA\|$ is large, σ is forced to be large, and the consequence is that the step size for solving (1.6a) becomes small and it is doomed that more outer iterations are needed, despite that the inner iterations can be avoided. In [20], it is shown that the best bound of σ is $0.75 \cdot r\|A^TA\|$ to ensure the convergence of (1.6), while the mentioned difficult remains if $\|A^TA\|$ is too large.

There is another algorithm that can be applied to the problem (1.1), and $\theta(x)$ and A can be decoupled in its implementation. More specifically, let us consider the Lagrangian function of (1.1) and its saddle-point reformulation, and then apply the primal-dual method proposed in [5]. With some tedious details skipped, the resulting iterative scheme can be written as

$$\begin{cases} x^{k+1} = \arg\min\big\{\theta(x) + \frac{r}{2}\|x - (x^k + \frac{1}{r}A^T\lambda^k)\|^2 \;\big|\; x \in \mathcal{X}\big\}. & (1.7a)\\ \lambda^{k+1} = \lambda^k - \frac{1}{s}\big(A(2x^{k+1} - x^k) - b\big), & (1.7b)\end{cases}$$

where $r > 0$ and $s > 0$ are parameters for the primal- and dual-variable subproblems, respectively. In (1.7a), $\theta(x)$ and A are also decoupled, and this subproblem is also reduced to $\text{Prox}_\theta^r(x)$ when $\mathcal{X} = \Re^n$. Nearly at the same time as [5], the primal-dual method proposed in [5] was explained as an application of the classic PPA in [22], and then this PPA explanation has been used to analyze the convergence for variants of the primal-dual method (1.7), as well as other first-order algorithms, in the literature, see, e.g. [2, 7, 29]. To ensure the convergence of (1.7), as analyzed in [5], the condition

$$rs > \|A^TA\| \tag{1.8}$$

is required. Following the PPA explanation in [22], the condition (1.8) is used to ensure the positive definiteness of the matrix that is used to definite the underlying PPA. We refer to, e.g. [5, 6, 22, 24] for some efficient applications of the primal-dual method (1.7) to some image reconstruction problems whose corresponding $\|A^TA\|$ is small. Therefore, despite that the objective function $\theta(x)$ and the coefficient matrix A are decoupled in notation, the subproblem (1.7a) is correlated implicitly with A because of the condition (1.8). Clearly, the same difficulties as those for implementing the PALM (1.5) should be tackled if $\|A^TA\|$ is too large.

Our main purpose is to balance the subproblems of the classic ALM (1.2) such that both subproblems could be easy for some applications. More specifically, let $r > 0$ and $\delta > 0$ be arbitrary constants; and define the positive definite matrix $H_0 \in \Re^{m\times n}$ as

$$H_0 := \big(\frac{1}{r}AA^T + \delta I_m\big). \tag{1.9}$$

Then, with $q_0^k := x^k + \frac{1}{r}A^T\lambda^k$, the classic ALM (1.2) for the problem (1.1) is balanced as

$$\text{(Balanced ALM)}\begin{cases} x^{k+1} = \arg\min\big\{\theta(x) + \frac{r}{2}\|x - q_0^k\|^2 \;\big|\; x \in \mathcal{X}\big\}, & (1.10a)\\ \lambda^{k+1} = \lambda^k - H_0^{-1}(A(2x^{k+1} - x^k) - b). & (1.10b)\end{cases}$$

Hence, for the model (1.1) with $\mathcal{X} = \Re^n$, the balanced ALM (1.10) is reduced to

$$\begin{cases} x^{k+1} = \text{Prox}_\theta^r(q_0^k), & (1.11a)\\ \lambda^{k+1} = \lambda^k - H_0^{-1}(A(2x^{k+1} - x^k) - b). & (1.11b)\end{cases}$$

In the x-subproblem (1.10a), it is easy to discern that $\theta(x)$ and A are decoupled while the parameter r is not restricted by any condition related to $\|A^T A\|$ explicitly or implicitly, and thus it could be as easy as estimating Prox_θ^r when $\mathcal{X} = \Re^n$. Moreover, the λ-subproblem (1.10b) is still very easy though it involves the matrix A and thus becomes slightly more difficult than (1.2b) or (1.7b), see Remark 2.1 for more details. In this sense, the subproblems of the classic ALM (1.2) are balanced in (1.10). Obviously, the balanced ALM (1.10) enjoys the proximity-induced feature while it can avoid possible tiny step sizes for the subproblems (1.10a) even when $\|A^T A\|$ is large. This is an essential difference of the balanced ALM (1.10) from the PALM (1.5) and the primal-dual method (1.7). We consider the balanced ALM (1.10) a necessary supplement to the classic ALM (1.2), especially for the case where $\text{Prox}_\theta^r(x)$ has a closed-form representation but $\|A^T A\|$ is large.

The rest of this paper is organized as follows. We state the model to be considered and generalize the balanced ALM (1.10) for this model in Section 2. Then, we conduct convergence analysis for the balanced ALM in Section 3. In Section 4, we extend our discussion to separable convex programming models and propose a splitting version of the balanced ALM. An alternative strategy for balancing is discussed in Section 5. In Section 6, from the PPA perspective, we briefly discuss how to further generalize the algorithms to be proposed in Sections 2-5. Finally, some conclusions are made in Section 7.

2 Model and algorithm

Note that the classic ALM (1.2) was proposed in the context of the canonical convex programming model with linear equality constraints (1.1). Despite that our initial aim is to consider the balanced ALM (1.10) for the model (1.1), the balanced ALM (1.10) does can be generalized to the more general convex programming model with both linear equality and inequality constraints. We thus present our work in a more general setting as below.

2.1 Model

Instead of (1.1), let us consider the following more general convex programming model with both linear equality and inequality constraints:

$$\min\{\theta(x) \mid Ax = b \text{ (or } \geq b), \, x \in \mathcal{X}\}, \tag{2.1}$$

in which the setting is same as (1.1) except that the linear inequality $Ax \geq b$ is also included. The solution set of (2.1) is assumed to be nonempty throughout our discussion. With the consideration of the more general model (2.1), the applicable range of the algorithm to be proposed is thus wider than that of the classic ALM (1.2).

2.2 Algorithm

Now, let us generalize the balanced ALM (1.10) to the model (2.1), and the name remains for simplicity. Recall that our main purpose is to balance the subproblems in the classic ALM (1.2).

Algorithm: the balanced ALM for (2.1)

Let $r > 0$ and $\delta > 0$ be arbitrary constants; H_0 be defined in (1.9). Denote

$$q_0^k := x^k + \frac{1}{r}A^T\lambda^k \ \text{ and } \ s_0^k := A(2x^{k+1} - x^k) - b.$$

Then, with (x^k, λ^k), the new iterate (x^{k+1}, λ^{k+1}) is generated via the following steps:

$$\begin{cases} x^{k+1} = \arg\min\big\{\theta(x) + \frac{r}{2}\|x - q_0^k\|^2 \mid x \in \mathcal{X}\big\}, & \text{(2.2a)} \\ \lambda^{k+1} = \arg\min\big\{\frac{1}{2}(\lambda - \lambda^k)H_0(\lambda - \lambda^k) + (s_0^k)^T\lambda \mid \lambda \in \Lambda\big\}. & \text{(2.2b)} \end{cases}$$

Remark 2.1. *It is easy to see that the balanced ALM (2.2) is reduced to the aforementioned (1.10) if the model (1.1) is considered. In particular, we have $\Lambda = \Re^m$ and thus the subproblem (2.2b) is reduced to finding λ^{k+1} such that*

$$H_0(\lambda - \lambda^k) = -s_0^k.$$

When $Ax \geq b$ is considered in (2.1)*, we have $\Lambda = \Re^m_+$. For this case, the subproblem* (2.2b) *is reduced to the standard quadratic programming with non-negative sign constraints*

$$\min\big\{\frac{1}{2}(\lambda - \lambda^k)H_0(\lambda - \lambda^k) + (s_0^k)^T\lambda \mid \lambda \in \Re^n_+\big\},$$

or equivalently, the linear complementarity problem

$$0 \leq \lambda \ \perp \ \{H_0(\lambda - \lambda^k) + s_0^k\} \geq 0.$$

Recall that the matrix H_0 defined in (1.9) is positive definite and it can be well conditioned with appropriate choices of r and δ. Hence, it is extremely easy to decompose H_0, e.g., by the Cholesky decomposition. Then, many benchmark solvers including the well-known Lemke algorithm and conjugate gradient method, can be found in various textbooks (e.g., [15, 28, 33]), monographs (e.g., [9]), and papers (e.g., [17, 18]).

Remark 2.2. *Recall that the balanced ALM (2.2) is featured by the fact that the function $\theta(x)$ and the coefficient matrix A are decoupled without any explicit or implicit condition related to A in (2.2a). Compared with the PALM (1.5) and the primal-dual method (1.7), the balanced ALM (2.2) also has two parameters, δ and r, whose only restriction is their sign. As we will show in Section 3, the only essential role of δ is to theoretically ensure the positive definiteness of the corresponding matrix H as defined in (3.7). Therefore, there is no particular motivation to tune δ for different applications, and it can be just fixed as a small value beforehand. For the parameter r, just as the same parameter in the classic ALM (1.2), there is full-extent flexibility to tune this parameter. Certainly, how to tune r depends on the specific model and dataset under discussion, whilst there is no generic and unified theory to determine the optimal choice for all cases.*

3 Convergence analysis

In this section, we prove the convergence of the balanced ALM (2.2), and estimate its worst-case convergence rate measured by the iteration complexity.

3.1 Variational inequality characterization of (2.1)

Following our previous works [22,23], our analysis will be conducted in the variational inequality (VI) context. We first derive the VI characterization for the optimality condition of the model (2.1). Let the Lagrangian function of the problem (2.1) be defined as

$$L(x,\lambda) = \theta(x) - \lambda^T(Ax-b), \tag{3.1}$$

with $\lambda \in \Re^m$ the Lagrange multiplier. Since both linear equality and inequality constraints are considered in (2.1), let us define

$$\Omega := \mathcal{X} \times \Lambda \quad \text{where} \quad \Lambda := \left\{ \begin{array}{ll} \Re^m, & \text{if } Ax = b, \\ \Re^m_+, & \text{if } Ax \geq b. \end{array} \right. \tag{3.2}$$

The pair $(x^*,\lambda^*) \in \Omega$ is called a saddle point of the Lagrangian function (3.1) if it satisfies the inequalities

$$L_{\lambda\in\Lambda}(x^*,\lambda) \leq L(x^*,\lambda^*) \leq L_{x\in\mathcal{X}}(x,\lambda^*).$$

Alternatively, we can write these inequalities as the following VIs:

$$\left\{ \begin{array}{lrl} x^* \in \mathcal{X}, & \theta(x) - \theta(x^*) + (x-x^*)^T(-A^T\lambda^*) \geq 0, & \forall\, x \in \mathcal{X}, \\ \lambda^* \in \Lambda, & (\lambda-\lambda^*)^T(Ax^*-b) \geq 0, & \forall\, \lambda \in \Lambda, \end{array} \right. \tag{3.3}$$

or in the compact format

$$w^* \in \Omega, \quad \theta(x) - \theta(x^*) + (w-w^*)^T F(w^*) \geq 0, \quad \forall\, u \in \Omega, \tag{3.4a}$$

where

$$w = \left(\begin{array}{c} x \\ \lambda \end{array} \right), \quad F(w) = \left(\begin{array}{c} -A^T\lambda \\ Ax-b \end{array} \right) \quad \text{and} \quad \Omega = \mathcal{X} \times \Lambda. \tag{3.4b}$$

Note that the operator F defined in (3.4b) is affine with a skew-symmetric matrix and thus we have

$$(w-\tilde{w})^T(F(w)-F(\tilde{w})) \equiv 0. \tag{3.5}$$

We also call (3.4) a monotone mixed variational inequality because the function θ is convex and the operator F has the property (3.5). We denote by Ω^* the solution set of the VI (3.4); it is also the set of the saddle points of the Lagrangian function (3.1).

3.2 Contraction

We need to show that the sequence generated by the balanced ALM (2.2) is contractive with respect to Ω^*, the solution set of the VI (3.4). This is the key property to ensure its convergence. Before that, let us recall a basic lemma whose proof is elementary and can be found in, e.g., [1].

Lemma 3.1. *Let $\mathcal{X} \subset \Re^n$ be a closed convex set, $\theta(x)$ and $f(x)$ be convex functions. If f is differentiable, and the solution set of the minimization problem*

$$\min\{\theta(x) + f(x) \,|\, x \in \mathcal{X}\}$$

is nonempty, then it holds that

$$x^* \in \arg\min\{\theta(x) + f(x) \,|\, x \in \mathcal{X}\} \tag{3.6a}$$

if and only if

$$x^* \in \mathcal{X}, \quad \theta(x) - \theta(x^*) + (x-x^*)^T \nabla f(x^*) \geq 0, \quad \forall\, x \in \mathcal{X}. \tag{3.6b}$$

To show the contraction property of the sequence generated by the balanced ALM (2.2), the first step is to fathom the difference of an iterate generate by the balanced ALM (2.2) from a solution point $w^* \in \Omega^*$. Recall the definition of H_0 in (1.9). Let us define

$$H = \begin{pmatrix} rI_n & A^T \\ A & \frac{1}{r}AA^T + \delta I_m \end{pmatrix} = \begin{pmatrix} rI_n & A^T \\ A & H_0 \end{pmatrix}. \tag{3.7}$$

Proposition 3.1. *The matrix H defined in (3.7) is positive definite.*

Proof. Notice that

$$H = \begin{pmatrix} rI_n & A^T \\ A & \frac{1}{r}AA^T \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \delta I_m \end{pmatrix} = \begin{pmatrix} \sqrt{r}I_n \\ \sqrt{\frac{1}{r}}A \end{pmatrix} \left(\sqrt{r}I_n, \sqrt{\tfrac{1}{r}}A^T\right) + \begin{pmatrix} 0 & 0 \\ 0 & \delta I_m \end{pmatrix},$$

for any $w = (x, \lambda) \neq 0$. Thus, we have

$$w^T H w = \left\|\sqrt{r}x + \sqrt{\tfrac{1}{r}}A^T\lambda\right\|^2 + \delta\|\lambda\|^2 > 0,$$

and therefore the matrix H is positive definite. □

In the following theorem, we will express the difference of an iterate generated by the balanced ALM (2.2) from a solution point $w^* \in \Omega^*$ in the context of VIs.

Theorem 3.1. *Let $\{w^k = (x^k, \lambda^k)\}$ be the sequence generated by the balanced ALM* (2.2) *and H be defined in (3.7). Then we have*

$$w^{k+1} \in \Omega, \;\; \theta(x) - \theta(x^{k+1}) + (w - w^{k+1})^T F(w^{k+1}) \geq (w - w^{k+1})^T H(w^k - w^{k+1}), \quad \forall\, w \in \Omega. \tag{3.8}$$

Proof. According to Lemma 3.1, the solution x^{k+1} of the subproblem (2.2a) can be characterized by the VI

$$x^{k+1} \in \mathcal{X}, \;\; \theta(x) - \theta(x^{k+1}) + (x - x^{k+1})^T\{-A^T\lambda^k + r(x^{k+1} - x^k)\} \geq 0, \;\; \forall\, x \in \mathcal{X}.$$

Then, for any unknown λ^{k+1}, we have

$$\begin{aligned} x^{k+1} \in \mathcal{X}, \;\; & \theta(x) - \theta(x^{k+1}) + (x - x^{k+1})^T(-A^T\lambda^{k+1}) \\ & \geq (x - x^{k+1})^T\{r(x^k - x^{k+1}) + A^T(\lambda^k - \lambda^{k+1})\}, \;\; \forall\, x \in \mathcal{X}. \end{aligned} \tag{3.9}$$

Similarly, because of Lemma 3.1, the solution λ^{k+1} of the subproblem (2.2b) can be characterized by the VI

$$\lambda^{k+1} \in \Lambda, \;\; (\lambda - \lambda^{k+1})^T\Big\{\Big(A[2x^{k+1} - x^k] - b\Big) + H_0(\lambda^{k+1} - \lambda^k)\Big\} \geq 0, \quad \forall\, \lambda \in \Lambda.$$

Recall the definition of H_0 in (1.9). We thus have

$$\begin{aligned} \lambda^{k+1} \in \Lambda, \;\; & (\lambda - \lambda^{k+1})^T(Ax^{k+1} - b) \\ & \geq (\lambda - \lambda^{k+1})^T\{(A(x^k - x^{k+1}) + \Big(\frac{1}{r}AA^T + \delta I_m\Big)(\lambda^k - \lambda^{k+1})\}, \;\; \forall\, \lambda \in \Lambda. \end{aligned} \tag{3.10}$$

Combining (3.9) and (3.10), and using the notation in (3.4), we obtain the assertion (3.8). □

In the following theorem, we will prove an important inequality which measures the difference of an iterate generated by the balanced ALM (2.2) from a solution point $w^* \in \Omega^*$ more explicitly by H-norm-induced distances. This inequality is also the basis of estimating the convergence rate measured by the iteration complexity for the balanced ALM (2.2).

Theorem 3.2. *Let $\{w^k = (x^k, \lambda^k)\}$ be the sequence generated by the balanced ALM* (2.2) *and H be defined in (3.7). Then we have*

$$\begin{aligned}&\theta(x) - \theta(x^{k+1}) + (w - w^{k+1})^T F(w) \\ &\geq \ \frac{1}{2}\big(\|w - w^{k+1}\|_H^2 - \|w - w^k\|_H^2\big) + \frac{1}{2}\|w^k - w^{k+1}\|_H^2, \ \ \forall w \in \Omega. \end{aligned} \tag{3.11}$$

Proof. It follows from (3.5) that

$$(w - w^{k+1})^T F(w^{k+1}) = (w - w^{k+1})^T F(w),$$

and thus the left-hand side of (3.8) equals

$$\theta(x) - \theta(x^{k+1}) + (w - w^{k+1})^T F(w).$$

Consequently, because of (3.8), we get

$$w^{k+1} \in \Omega, \quad \theta(x) - \theta(\tilde{x}^k) + (w - w^{k+1})^T F(w) \geq (w - w^{k+1})^T H(w^k - w^{k+1}), \quad \forall w \in \Omega. \tag{3.12}$$

Applying the identity

$$b^T H(b - a) = \frac{1}{2}\{\|b\|_H^2 - \|a\|_H^2\} + \frac{1}{2}\|a - b\|_H^2$$

to the right-hand side of (3.12) with $a = w - w^k$ and $b = w - w^{k+1}$, we thus obtain

$$(w - w^{k+1})^T H(w^k - w^{k+1}) \ = \ \frac{1}{2}\big(\|w - w^{k+1}\|_H^2 - \|w - w^k\|_H^2\big) + \frac{1}{2}\|w^k - w^{k+1}\|_H^2. \tag{3.13}$$

Substituting (3.13) into the right-hand side of (3.12), we prove the assertion (3.11). □

Now, with Theorems 3.1 and 3.2, the contraction property of the sequence generated by the balanced ALM (2.2) with respect to Ω^* can be proved.

Theorem 3.3. *Let $\{w^k = (x^k, \lambda^k)\}$ be the sequence generated by the balanced ALM* (2.2) *and H be defined in (3.7). Then we have*

$$\|w^{k+1} - w^*\|_H^2 \leq \|w^k - w^*\|_H^2 - \|w^k - w^{k+1}\|_H^2, \ \ \forall w^* \in \Omega^*. \tag{3.14}$$

Proof. Setting w in (3.11) as any fixed $w^* \in \Omega^*$, we get

$$\begin{aligned}&\|w^k - w^*\|_H^2 - \|w^{k+1} - w^*\|_H^2 - \|w^k - w^{k+1}\|_H^2 \\ &\quad \geq \ 2\big\{\theta(x^{k+1}) - \theta(x^*) + (w^{k+1} - w^*)^T F(w^*)\big\}, \ \ \forall w^* \in \Omega^*.\end{aligned}$$

Since $w^* \in \Omega^*$ and $w^{k+1} \in \Omega$, according to (3.4), the right-hand side of the last inequality is non-negative. Thus, the assertion of this theorem follows directly. □

3.3 Convergence

With the contraction property established in Theorem 3.3, it is easy to prove the convergence of the sequence $\{w^k\}$ generated by the balanced ALM (2.2).

Theorem 3.4. *Let $\{w^k = (x^k, \lambda^k)\}$ be the sequence generated by the balanced ALM* (2.2) *and H be defined in (3.7). Then, the sequence $\{w^k\}$ converges to some $w^\infty \in \Omega^*$.*

Proof. First of all, it follows from (3.14) that the sequence $\{w^k\}$ is bounded and

$$\lim_{k\to\infty} \|w^k - w^{k+1}\|_H^2 = 0. \tag{3.15}$$

Let w^∞ be a cluster point of $\{w^k\}$ and $\{w^{k_j}\}$ be a subsequence converging to w^∞. It follows from (3.8) that

$$w^{k_j} \in \Omega, \;\; \theta(x) - \theta(x^{k_j}) + (w - w^{k_j})^T F(w^{k_j}) \ge (w - w^{k_j})^T H(w^{k_j-1} - w^{k_j}), \quad \forall\, w \in \Omega.$$

Since the matrix H is positive definite, it follows from (3.15) and the continuity of $\theta(x)$ and $F(w)$ that

$$w^\infty \in \Omega, \;\; \theta(x) - \theta(x^\infty) + (w - w^\infty)^T F(w^\infty) \ge 0, \quad \forall\, w \in \Omega.$$

This VI above indicates that w^∞ is a solution point of (3.4). Finally, because of (3.14), we have

$$\|w^{k+1} - w^\infty\|_H^2 \le \|w^k - w^\infty\|_H^2,$$

and thus $\{w^k\}$ converges to w^∞. The proof is complete. □

3.4 Convergence rate

Following the VI-based technique established in our earlier work [23], we can estimate the worst-case $O(1/t)$ convergence rate measured by the iteration complexity for the balanced ALM (2.2) where t is the iteration counter.

Let us recall some necessary details which can also be found in [23]. If $\tilde w$ is a solution point of the VI (3.4), then we have

$$\tilde w \in \Omega, \quad \theta(x) - \theta(\tilde x) + (w - \tilde w)^T F(\tilde w) \ge 0, \quad \forall\, w \in \Omega.$$

Because of (3.5), $\tilde w$ also satisfies

$$\tilde w \in \Omega, \quad \theta(x) - \theta(\tilde x) + (w - \tilde w)^T F(w) \ge 0, \quad \forall\, w \in \Omega.$$

Thus, for given $\epsilon > 0$, $\tilde w \in \Omega$ is called an ϵ-approximate solution of VI (3.4) if it satisfies

$$\tilde w \in \Omega, \;\; \theta(x) - \theta(\tilde x) + (w - \tilde w)^T F(w) \ge -\epsilon, \;\; \forall\; w \in \mathcal{D}_{(\tilde w)}, \tag{3.16}$$

where

$$\mathcal{D}_{(\tilde w)} = \{w \in \Omega \,|\, \|w - \tilde w\| \le 1\}.$$

Thus, to establish the worst-case $O(1/t)$ convergence rate for the balanced ALM (2.2), we need to show that, for given $\epsilon > 0$, after t iterations, we can find $\tilde w \in \Omega$, such that

$$\tilde w \in \Omega, \quad \text{and} \quad \sup_{w\in\mathcal{D}_{(\tilde w)}} \big\{\theta(\tilde x) - \theta(x) + (\tilde w - w)^T F(w)\big\} \le \epsilon = O(1/t). \tag{3.17}$$

We present this result in the following theorem.

Theorem 3.5. *Let $\{w^k = (x^k, \lambda^k)\}$ be the sequence generated by the balanced ALM (2.2) and H be defined in (3.7). For any integer number $t > 0$, if we define*

$$\tilde w_t := \frac{1}{t+1}\sum_{k=0}^{t} w^{k+1}, \tag{3.18}$$

then we have

$$\tilde w_t \in \Omega, \quad \theta(\tilde x_t) - \theta(x) + (\tilde w_t - w)^T F(w) \le \frac{1}{2(t+1)}\|w - w^0\|_H^2, \quad \forall w \in \Omega. \tag{3.19}$$

Proof. First, it follows from (3.11) that, for all $k \geq 0$, we have

$$w^{k+1} \in \Omega, \quad \theta(x) - \theta(x^{k+1}) + (w - w^{k+1})^T F(w) + \frac{1}{2}\|w - w^k\|_H^2 \geq \frac{1}{2}\|w - w^{k+1}\|_H^2, \ \forall w \in \Omega. \quad (3.20)$$

Summarizing the inequalities (3.20) over $k = 0, 1, \ldots, t$, we obtain

$$(t+1)\theta(x) - \sum_{k=0}^{t} \theta(x^{k+1}) + \Big((t+1)w - \sum_{k=0}^{t} w^{k+1}\Big)^T F(w) + \frac{1}{2}\|w - w^0\|_H^2 \geq 0, \quad \forall w \in \Omega.$$

It follows from (3.18) that

$$\frac{1}{t+1}\sum_{k=0}^{t} \theta(x^{k+1}) - \theta(x) + (\tilde{w}_t - w)^T F(w) \leq \frac{1}{2(t+1)}\|w - w^0\|_H^2, \quad \forall w \in \Omega. \quad (3.21)$$

Note that $\tilde{w}_t$ defined in (3.18) is a convex combination of all iterates w^k for $k = 0, \cdots, t$, and $\theta(x)$ is convex. We thus have

$$\tilde{x}_t = \frac{1}{t+1}\sum_{k=0}^{t} x^{k+1},$$

and also

$$\theta(\tilde{x}_t) \leq \frac{1}{t+1}\sum_{k=0}^{t} \theta(x^{k+1}).$$

Substituting it into (3.21), the assertion (3.19) of this theorem follows directly. □

Then, because of (3.17), the inequality (3.19) indicates that $\tilde{w}_t$ defined in (3.18), which is the average of the first t iterates generated by the balanced ALM (2.2), is an approximate solution of the VI (3.4) with an accuracy of $O(1/t)$. Hence, the worst-case $O(1/t)$ convergence rate measured by the iteration complexity is established for the balanced ALM (2.2) in the ergodic sense.

4 Splitting versions of the balanced ALM (2.2) for separable convex programming

The classic ALM (1.2) plays an extremely influential role in solving various separable cases of the generic model (1.1) when the objective function of such a model can be represented as the sum of multiple functions without coupled variables. For these separable models, the classic ALM (1.2) has been adapted into various splitting versions by decomposing the primeval x-subproblem (1.2a) into smaller ones. These splitting versions take advantage of the separable structure in the model more effectively; the decomposed subproblems are usually easier in the sense that each of them only needs to tackle one function component. For various applications including the mentioned sparsity- and low-rank-promoted ones in data science domains, splitting versions of the ALM (1.2) may generate subproblems that are easy enough to have closed-form solutions. Among various splitting versions of the classic ALM (1.2), the most popular one is probably the mentioned ADMM in [13], which suggests splitting the x-subproblem (1.2a) into two sequentially when the model (1.1) has a two-block separable structure.

In this section, in parallel with the successful legacy of the classic ALM (1.2) and its splitting versions, we also discuss how to design splitting versions for the balanced ALM (2.2) when the model (2.1) is separable. For succinctness and without ambiguity, we reuse some letters and notation as those in Sections 2 and 3.

4.1 Model

Let us consider the separable convex programming model with both linear equality and inequality constraints

$$\min\Big\{\sum_{i=1}^{p}\theta_i(x_i)\mid\sum_{i=1}^{p}A_ix_i=b\ (\text{or}\geq b),\ \ x_i\in\mathcal{X}_i\Big\},\tag{4.1}$$

where $\theta_i:\Re^{n_i}\to\Re$, $i=1,\ldots,p$, are closed proper convex functions and they are not necessarily smooth; $\mathcal{X}_i\subseteq\Re^{n_i}$, $i=1,\ldots,p$, are closed convex sets; $A_i\in\Re^{m\times n_i}$, $i=1,\ldots,p$, are given matrices; and $b\in\Re^m$ is a given vector. The model (4.1) can be regarded as an extension of the model (2.1) from $p=1$ to $p\geq 1$. Let us only consider the multiple-block separable case with $p\geq 2$ for (4.1). Similarly as (3.2), we reuse the letters and define

$$\Omega=\prod_{i=1}^{p}\mathcal{X}_i\times\Lambda\qquad\text{where}\qquad\Lambda=\begin{cases}\Re^m, & \text{if }\sum_{i=1}^{p}A_ix_i=b,\\ \Re^m_+, & \text{if }\sum_{i=1}^{p}A_ix_i\geq b.\end{cases}\tag{4.2}$$

4.2 Algorithm

Now, we extend the balanced ALM (2.2) to the multiple-block separable convex programming model (4.1) and present a splitting version of (2.2) below.

Algorithm: A splitting version of the balanced ALM (2.2) for (4.1)

Let $r_i>0$ for $i=1,2,\cdots,p$, and $\delta>0$ be arbitrary constants. Define

$$H_p=\sum_{i=1}^{p}\frac{1}{r_i}A_iA_i^T+\delta I_m,\tag{4.3}$$

$$q_i^k:=x_i^k+\frac{1}{r_i}A^T\lambda^k,\ \text{for }i=1,2,\cdots,p;\ \text{and}\ s^k=\sum_{i=1}^{p}A_i(2x_i^{k+1}-x_i^k)-b.$$

Then, with $w^k=(x_1^k,x_2^k,\cdots,x_p^k,\lambda^k)$, the new iterate $w^{k+1}=(x_1^{k+1},x_2^{k+1},\cdots,x_p^{k+1},\lambda^{k+1})$ is generated via the following steps:

$$\begin{cases}x_i^{k+1}\in\arg\min\big\{\theta_i(x_i)+\frac{r_i}{2}\|x_i-q_i^k\|^2\mid x_i\in\mathcal{X}_i\big\},i=1,2,\cdots,p; & \text{(4.4a)}\\ \lambda^{k+1}=\arg\min\left\{\frac{1}{2}(\lambda-\lambda^k)^TH_p(\lambda-\lambda^k)+(s^k)^T\lambda\mid\lambda\in\Lambda\right\}. & \text{(4.4b)}\end{cases}$$

Remark 4.1. *The subproblems in (4.4) are of the same structure as those in (2.2). For the x_i-subproblem (4.4a), the function $\theta_i(x_i)$ and the coefficient A_i are decoupled without any explicit or implicit condition related to A_i, and thus it is also reduced to estimating the proximity operator of $\theta_i(x_i)$ when $\mathcal{X}_i=\Re^{n_i}$. In addition, the λ-subproblem (4.4b) is a positive definite system of linear equations or a standard quadratic programming with non-negative sign constraints. Note that all r_i's have no other restriction than the sign requirement. Hence, the algorithm (4.4) keeps all features of the balanced ALM (2.2).*

Remark 4.2. *For generality, we consider different r_i for different x_i-subproblems. They can be identical for simplicity. Similarly as the balanced ALM (2.2), to implement the algorithm (4.4), empirically we can fix δ as a small positive constant throughout.*

4.3 Convergence analysis

In this subsection, we follow the analysis in Section 3 and prove the convergence of the splitting version of the balanced ALM (4.4).

4.3.1 Variational inequality characterization of (4.1)

For convergence analysis purpose, we also need the VI characterization for the optimality condition of the model (4.1). Let $\lambda \in \Re^m$ be the Lagrange multiplier of (4.1) and the Lagrangian function of the problem (4.1) be defined as

$$L(x_1, \ldots, x_p, \lambda) = \sum_{i=1}^{p} \theta_i(x_i) - \lambda^T \Big(\sum_{i=1}^{p} A_i x_i - b \Big). \tag{4.5}$$

Similarly as Section 3.1, we reuse the letters and know that finding a saddle point of $L(x_1, \ldots, x_p, \lambda)$ can be written as the following VI:

$$w^* \in \Omega, \quad \theta(x) - \theta(x^*) + (w - w^*)^T F(w^*) \geq 0, \quad \forall\, w \in \Omega, \tag{4.6a}$$

where

$$w = \begin{pmatrix} x_1 \\ \vdots \\ x_p \\ \lambda \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \quad \theta(x) = \sum_{i=1}^{p} \theta_i(x_i), \quad F(w) = \begin{pmatrix} -A_1^T \lambda \\ \vdots \\ -A_p^T \lambda \\ \sum_{i=1}^{p} A_i x_i - b \end{pmatrix}, \tag{4.6b}$$

and Ω is defined in (4.2). Again, we denote by Ω^* the solution set of the VI (4.6).

4.3.2 Convergence

Let us recall the proofs in Section 3 for the convergence of the balanced ALM (2.2). It is easy to see that the crucial step is to identify the difference between an iterate and a solution point by the inequality (3.8) in Theorem 3.1, in which the matrix H should be positive definite as proved in Proposition 3.1 so that the difference can be measured by distances defined by the H-norm. After Proposition 3.1 and Theorem 3.1 are proved, the remaining part of the proof for the convergence as well as the worst-case convergence rate is subroutine. Hence, to prove the convergence of the splitting version of the balanced ALM (4.4), we only need to prove an inequality similar as (3.8) in which the accompanying matrix is also positive definite.

Proposition 4.1. *Let $r_i > 0$ for $i = 1, 2, \cdots, p$, and $\delta > 0$ be arbitrary constants. The matrix defined as*

$$H = \begin{pmatrix} r_1 I_{n_1} & 0 & \cdots & 0 & A_1^T \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & r_p I_{n_p} & A_p^T \\ A_1 & \cdots & \cdots & A_p & \displaystyle\sum_{i=1}^{p} \frac{1}{r_i} A_i A_i^T + \delta I_m \end{pmatrix} \tag{4.7}$$

is positive definite.

Proof. Note that

$$H = \sum_{i=1}^{p} H_i + \begin{pmatrix} 0 & 0 \\ 0 & \delta I_m \end{pmatrix},$$

where

$$H_i = \begin{pmatrix} r_i I_{n_i} & A_i^T \\ & \\ A_i & \frac{1}{r_i} A_i A_i^T \end{pmatrix} = \begin{pmatrix} \vdots \\ \sqrt{r_i} I_{n_i} \\ \vdots \\ \sqrt{\frac{1}{r_i}} A_i \end{pmatrix} \begin{pmatrix} \cdots & \sqrt{r_i} I_{n_i} & \cdots & \sqrt{\frac{1}{r_i}} A_i^T \end{pmatrix}.$$

For any $w = (x_1, \ldots, x_p, \lambda) \neq 0$, we have

$$w^T H w = \sum_{i=1}^{p} \left\| \sqrt{r_i} x_i + \sqrt{\frac{1}{r_i}} A_i^T \lambda \right\|^2 + \delta \|\lambda\|^2 > 0.$$

Hence, the matrix H is positive definite. □

Theorem 4.1. *Let $\{w^k = (x_1^k, \cdots, x_p^k, \lambda^k)\}$ be the sequence generated by the balanced ALM (4.4) and H be defined in (4.7). Then, we have*

$$w^{k+1} \in \Omega, \;\; \theta(x) - \theta(x^{k+1}) + (w - w^{k+1})^T F(w^{k+1}) \geq (w - w^{k+1})^T H (w^k - w^{k+1}), \quad \forall\, w \in \Omega. \tag{4.8}$$

Proof. According to Lemma 3.1, for $i = 1, 2, \ldots, p$, we have

$$x_i^{k+1} \in \mathcal{X}_i, \quad \theta_i(x_i) - \theta_i(\tilde{x}_i^k) + (x - x_i^{k+1})^T \{-A_i^T \lambda^k + r_i(x_i^{k+1} - x_i^k)\} \geq 0, \;\; \forall\, x_i \in \mathcal{X}_i.$$

Then, for any unknown λ^{k+1}, we have

$$\begin{aligned} x_i^{k+1} \in \mathcal{X}_i, \;\; & \theta_i(x_i) - \theta_i(x_i^{k+1}) + (x_i - x_i^{k+1})^T(-A^T \lambda^{k+1}) \\ & \geq (x_i - x_i^{k+1})^T \big\{ r_i(x_i^k - x_i^{k+1}) + A^T(\lambda^k - \lambda^{k+1}) \big\}, \;\; \forall\, x_i \in \mathcal{X}_i. \end{aligned} \tag{4.9}$$

Also because of Lemma 3.1, λ^{k+1} generated by (4.4b) is characterized by the VI

$$\lambda^{k+1} \in \Lambda, \;\; (\lambda - \lambda^{k+1})^T \Big\{ \Big(\sum_{i=1}^{p} A_i [2x_i^{k+1} - x_i^k] - b \Big) + \Big(\sum_{i=1}^{p} \frac{1}{r_i} A_i A_i^T + \delta I_m \Big)(\lambda^{k+1} - \lambda^k) \Big\} \geq 0, \quad \forall\, \lambda \in \Lambda.$$

It can be rewritten as

$$\begin{aligned} \lambda^{k+1} \in \Lambda, \;\; & (\lambda - \lambda^{k+1})^T \Big(\sum_{i=1}^{p} A_i x_i^{k+1} - b \Big) \\ & \geq (\lambda - \lambda^{k+1})^T \Big\{ \sum_{i=1}^{p} A_i (x_i^k - x_i^{k+1}) + \Big(\sum_{i=1}^{p} \frac{1}{r_i} A_i A_i^T + \delta I_m \Big)(\lambda^k - \lambda^{k+1}) \Big\}, \forall\, \lambda \in \Lambda \end{aligned} \tag{4.10}$$

Combining (4.9) and (4.10), and using the notation in (4.6), we prove the assertion (4.8). □

As mentioned, based on Proposition 4.1 and Theorem 4.1, similar conclusions as Theorems 3.2-3.5 can be trivially proved. Thus, convergence results similar as those in Section 3 can be obtained for the splitting version of the balanced ALM (4.4); we omit the details for succinctness.

5 An alternative strategy for balancing

The balanced ALM (2.2) can be generalized to the splitting version (4.4) if the model under discussion is changed from the one-block case (2.1) to the multiple-block case (4.1). There are other ways for the generalization, in addition to the technique introduced in Section 4. In (4.4), we see that each of the x_i-subproblems does not involve any quadratic term in form of $\frac{r_i}{2}\|A_i x_i - q_i^k\|^2$ so that it can be reduced to estimating the proximity operator of $\theta_i(x_i)$ when $\mathcal{X}_i = \Re^{n_i}$. In this sense, all such x_i-subproblems are preferred when it is easy to estimate the proximity operator of $\theta_i(x_i)$. On the other hand, all A_i's are aggregated in the λ-subproblem (4.4b) because of the matrix H_p defined in (4.3). For some cases where some or all $\|A_i^T A_i\|$ are large (or, some or all A_i's are ill-conditioned), it is preferred to consider alleviating the quadratic programming problem (4.4b) by removing such $A_i^T A_i$ from H_p. Hence, from methodological point of view, it is also interesting to ask if we can keep terms in form of $\frac{r_i}{2}\|A_i x_i - q_i^k\|^2$ for some x_i-subproblems (4.4a), and meanwhile remove the corresponding $A_i A_i^T$ from the matrix H_p in (4.3) so that the λ-subproblem (4.4b) becomes easier. Accordingly, we propose to revise the splitting version of the balanced ALM (4.4) such that some x_i-subproblems are in form of

$$x_i^{k+1} \in \arg\min\bigl\{\theta_i(x_i) + \frac{r_i}{2}\|A_i x_i - q_i^k\|^2 \mid x_i \in \mathcal{X}_i\bigr\},$$

with q_i^k a certain constant vector, and the corresponding $A_i A_i^T$ is excluded in the λ-subproblem (4.4b). This idea provides an alternative strategy for balancing the generated subproblems, and it enables a user to determine how to balance the difficulty of subproblems in accordance with the specific functions θ_i's, coefficient matrices A_i's, and sets $\mathcal{X}_i$'s for a given application.

5.1 Model

For succinctness of notation, let us just take the special case of (4.1) with $p = 2$ and only linear equality constraints to illustrate our idea:

$$\min\{\theta_1(x_1) + \theta_2(x_2) \mid A_1 x_1 + A_2 x_2 = b,\ x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2\}. \tag{5.1}$$

Again, without ambiguity, some letters and notation are reused.

5.2 Algorithm

An alternative splitting version of the balanced ALM (2.2) for the specific model (5.1) can be presented as below.

Algorithm: An alternative splitting version of the balanced ALM for (5.1)

Let $r > 0$, $s > 0$ and $\delta > 0$ be arbitrary constants. Define

$$H_2 = \frac{1}{s}A_2 A_2^T + (\frac{1}{r} + \delta)I_m, \tag{5.2}$$

$$q_2^k := x_2^k + \frac{1}{s}A_2^T\lambda^k \text{ and } s_2^k = A_1(2x_1^{k+1} - x_1^k) + A_2(2x_2^{k+1} - x_2^k) - b.$$

Then, with $w^k = (x_1^k, x_2^k, \lambda^k)$, the new iterate $w^{k+1} = (x_1^{k+1}, x_2^{k+1}, \lambda^{k+1})$ is generated via the following steps:

$$\left\{\begin{array}{ll} x_1^{k+1} = \arg\min\bigl\{\theta_1(x_1) - x_1^T A_1^T \lambda^k + \frac{r}{2}\|A_1(x_1 - x_1^k)\|^2 + \frac{\delta}{2}\|x_1 - x_1^k\|^2 \bigm| x_1 \in \mathcal{X}_1\bigr\}, & (5.3a)\\ x_2^{k+1} = \arg\min\bigl\{\theta_2(x_2) + \frac{s}{2}\|x_2 - q_2^k\|^2 \bigm| x_2 \in \mathcal{X}_2\bigr\}, & (5.3b)\\ \lambda^{k+1} = \arg\min\Bigl\{\frac{1}{2}(\lambda - \lambda^k)H_2(\lambda - \lambda_k) + (s_2^k)^T\lambda \mid \lambda \in \Lambda\Bigr\}. & (5.3c) \end{array}\right.$$

Remark 5.1. *In the algorithm (5.3), we see that only the x_2-subproblem (5.3b) can be reduced to estimating the proximity operator of θ_2 if $\mathcal{X}_2 = \Re^{n_2}$, while the x_1-subproblem (5.3a) is not proximity-induced because the term $\|A_1(x_1 - x_1^k)\|^2$ is kept. As a balance, the matrix H_2 defined in (5.2) which determines the quadratic programming problem (5.3c) does not involve A_1. In this sense, the x_i-subproblems and the λ-subproblem are balanced in another way. For the generic model (4.1) with $p > 2$, the splitter version of the balanced ALM (4.4) can be revised in the sense that some of its x_i-subproblems are flexibly chosen to keep the terms $\|A_i(x_i - x_i^k)\|^2$ whilst the quadratic term determining the λ^k-subproblem does not involve the corresponding A_i's. Thus, different algorithms with different balanced subproblems can be designed analogously. The algorithm (5.3) is just the simplest illustration with $p = 2$ for this philosophy.*

5.3 Convergence results

As mentioned, to prove the convergence of the algorithm (5.3), we just need to prove an inequality similar as (3.8) in Theorem 3.1 and show that the accompanying matrix is positive definite.

Proposition 5.1. *Let $r > 0$, $s > 0$, and $\delta > 0$ be arbitrary constants. Then, the matrix defined as*

$$H = \begin{pmatrix} rA_1^TA_1 + \delta I_{n_1} & 0 & A_1^T \\ 0 & sI_{n_2} & A_2^T \\ A_1 & A_2 & \frac{1}{s}A_2A_2^T + (\frac{1}{r} + \delta)I_m \end{pmatrix} \tag{5.4}$$

is positive definite.

Proof. Note that

$$H = \begin{pmatrix} rA_1^TA_1 + \delta I_{n_1} & 0 & A_1^T \\ 0 & 0 & 0 \\ A & 0 & \frac{1}{r}I_m \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ & sI_{n_2} & A_2^T \\ 0 & A_2 & \frac{1}{s}A_2A_2^T \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \delta I_m \end{pmatrix}.$$

For any $w = (x, y, \lambda) \neq 0$, we have

$$w^THw = \Big(\Big\|\sqrt{r}A_1x + \sqrt{\tfrac{1}{r}}\lambda\Big\|^2 + \delta\|x\|^2\Big) + \Big\|\sqrt{s}y + \sqrt{\tfrac{1}{s}}A_2^T\lambda\Big\|^2 + \delta\|\lambda\|^2 > 0.$$

Thus, the matrix H is positive definite. □

Theorem 5.1. *Let $\{w^k = (x_1^k, x_2^k, \lambda^k)\}$ be the sequence generated by the algorithm (5.3) and H be defined in (5.4). Then, we have*

$$w^{k+1} \in \Omega, \;\; \theta(u) - \theta(u^{k+1}) + (w - w^{k+1})^T F(w^{k+1}) \geq (w - w^{k+1})^T H(w^k - w^{k+1}), \;\; \forall\, w \in \Omega. \tag{5.5}$$

Proof. According to Lemma 3.1, x^{k+1} generated by (5.3a) is characterized by the VI

$$x_1^{k+1} \in \mathcal{X}_1, \;\; \theta_1(x_1) - \theta_1(x_1^{k+1}) + (x_1 - x_1^{k+1})^T\big\{-A_1^T\lambda^k + (rA_1^TA_1 + \delta)(x_1^{k+1} - x_1^k)\big\} \geq 0, \;\; \forall\, x_1 \in \mathcal{X}_1.$$

Then, for any unknown λ^{k+1}, we have

$$\begin{aligned} x_1^{k+1} \in \mathcal{X}_1, \;\; & \theta_1(x_1) - \theta_2(x_1^{k+1}) + (x_1 - x_1^{k+1})^T(-A_1^T\lambda^{k+1}) \\ & \geq (x_1 - x_1^{k+1})^T\big\{(rA_1^TA_1 + \delta)(x_1^k - x_1^{k+1}) + A_1^T(\lambda^k - \lambda^{k+1})\big\}, \;\; \forall\, x_1 \in \mathcal{X}_1. \end{aligned} \tag{5.6}$$

Analogously, it follows from Lemma 3.1 that x_2^{k+1} generated by (5.3b) can be characterized by the VI

$$x_2^{k+1} \in \mathcal{X}_2, \;\; \theta_2(x_2) - \theta_2(x_2^{k+1}) + (x_2 - x_2^{k+1})^T\big\{-A_2^T\lambda^k + s(x_2^{k+1} - x_2^k)\big\} \geq 0, \;\; \forall\, x \in \mathcal{X}_2.$$

Then, for any unknown λ^{k+1}, we have

$$\begin{aligned} x_2^{k+1} \in \mathcal{X}_2, \;\; & \theta_2(x_2) - \theta_2(x_2^{k+1}) + (x_2 - x_2^{k+1})^T(-A_2^T\lambda^{k+1}) \\ & \ge (x_2 - x_2^{k+1})^T\big\{s(x_2^k - x_2^{k+1}) + A_2^T(\lambda^k - \lambda^{k+1})\big\}, \;\; \forall\, x_2 \in \mathcal{X}_2. \end{aligned} \tag{5.7}$$

Similarly, according to Lemma 3.1, λ^{k+1} generated by (5.3c) is characterized by the VI: Finding $\lambda^{k+1} \in \Lambda$ such that

$$(\lambda - \lambda^{k+1})^T\Big\{\Big(A_1[2x_1^{k+1} - x_1^k] + A_2[2x_2^{k+1} - x_2^k] - b\Big) + \Big(\frac{1}{s}A_2A_2^T + (\frac{1}{r} + \delta)I_m\Big)(\lambda^{k+1} - \lambda^k)\Big\} \ge 0, \;\; \forall\, \lambda \in \Lambda.$$

It can be rewritten as

$$\begin{aligned} & \lambda^{k+1} \in \Lambda, \quad (\lambda - \lambda^{k+1})^T(A_1x_1^{k+1} + A_2x_2^{k+1} - b) \\ & \ge (\lambda - \lambda^{k+1})^T\Big\{A_1(x_1^k - x_1^{k+1}) + A_2(x_2^k - x_2^{k+1}) + \Big(\frac{1}{s}A_2A_2^T + (\frac{1}{r} + \delta)I_m\Big)(\lambda^k - \lambda^{k+1})\Big\}, \end{aligned} \tag{5.8}$$

for all $\lambda \in \Lambda$. Combining (5.6), (5.7) and (5.8), and using the notation in (3.4), we get the following assertion. □

5.4 Comparison with linearized versions of the ADMM

It is interesting to compare the proposed algorithm (5.3) with the well-known linearized versions of the ADMM. For the model (5.1), the original ADMM scheme reads as

$$\begin{cases} x_1^{k+1} \in \arg\min\big\{\theta_1(x_1) - x_1^TA^T\lambda^k + \frac{r}{2}\|A_1x_1 + A_2x_2^k - b\|^2 \;\big|\; x_1 \in \mathcal{X}_1\big\}, & (5.9\text{a}) \\ x_2^{k+1} \in \arg\min\big\{\theta_2(x_2) - x_2^TA_2^T\lambda^k + \frac{r}{2}\|A_1x_1^{k+1} + A_2x_2 - b\|^2 \;\big|\; x_2 \in \mathcal{X}_2\big\}, & (5.9\text{b}) \\ \lambda^{k+1} = \lambda^k - r(A_1x_1^{k+1} + A_2x_2^{k+1} - b), & (5.9\text{c}) \end{cases}$$

in which $r > 0$ is the penalty parameter and $\lambda \in \Re^m$ is the Lagrange multiplier. The first proximal version of the ADMM (PADMM) which suggests regularizing both the x_1- and x_2-subproblems in (5.9) with generic proximal terms was proposed in [19] (see also [10] for a special case). For simplicity, let us assume that the x_1-subproblem (5.9a) is easy but the x_2-subproblem (5.9b) is difficult. Then, the PADMM in [19] can be written as

$$\begin{cases} x_1^{k+1} \in \arg\min\big\{\theta_1(x_1) - x_1^TA_1^T\lambda^k + \frac{r}{2}\|A_1x_1 + A_2x_2^k - b\|^2 \;\big|\; x_1 \in \mathcal{X}_1\big\}, & (5.10\text{a}) \\ x_2^{k+1} \in \arg\min\big\{\theta_2(x_2) - x_2^TA_2^T\lambda^k + \frac{r}{2}\|A_1x_1^{k+1} + A_2x_2 - b\|^2 + \frac{1}{2}\|x_2 - x_2^k\|_G^2 \;\big|\; x_2 \in \mathcal{X}_2\big\}, & (5.10\text{b}) \\ \lambda^{k+1} = \lambda^k - r(A_1x_1^{k+1} + A_2x_2^{k+1} - b), & (5.10\text{c}) \end{cases}$$

in which $G \in \Re^{n_2 \times n_2}$ is a positive definite matrix. Because of the same reason as mentioned for (1.5), it is interesting to consider "linearizing" the quadratic term "$\frac{r}{2}\|A_1x_1^{k+1} + A_2x_2 - b\|^2$" and thus alleviating the subproblem (5.10b) as estimating the proximity operator of $\theta_2(x_2)$ when $\mathcal{X}_2 = \Re^{n_2}$. Similar as (1.6), this can be done by choosing $G := sI_{n_2} - rA_2^TA_2$ in (5.10b). As well discussed in the literature, e.g., [11, 21, 26, 34, 35], for various applications arising in image processing, statistical learning, and others, the condition

$$s > r\|A_2^TA_2\| \tag{5.11}$$

is required to ensure the positive definiteness of G and thus the convergence of (5.10). Recently, the condition (5.11) is further optimally improved in [21] as $s > 0.75 \cdot r\|A_2^TA_2\|$. Similarly as (1.5a) and (1.7a), though $\theta_2(x_2)$ and A_2 are decoupled in notation if $G := sI_{n_2} - rA_2^TA_2$ in (5.10b), the subproblem (5.10b) is correlated implicitly with A_2 via the condition (5.11) or its improved one in [21]. Hence, efficiency of all existing linearized versions of the ADMM is severely affected if $\|A_2^TA_2\|$ is large. In this sense, the algorithm (5.3) improves existing linearized versions of the ADMM in the sense that the x_2-subproblem (5.3b) can be reduced to estimating the proximity operator of θ_2 if $\mathcal{X}_2 = \Re^{n_2}$, while it is not affected by $\|A_2^TA_2\|$ and thus possible tiny step sizes could be avoided even if $\|A_2^TA_2\|$ is large.

6 More generalized versions

In the preceding sections, the balanced ALM (2.2) is proposed for the generic model (2.1), and then its splitting versions (4.4) and (5.3) are studied for the separable models (4.1) and (5.1), respectively. As mentioned, it was shown in [32] that the classic ALM (1.2) is an application of the PPA proposed in [27]. In view of the generalized version of the PPA studied in [16], all the proposed algorithms (2.2), (4.4) and (5.3) can be further generalized. For instance, the balanced ALM (2.2) can be generalized as

$$\begin{cases} \tilde{x}^k = \arg\min\big\{\theta(x) + \dfrac{r}{2}\|x - q_0^k\|^2 \mid x \in \mathcal{X}\big\}, & (6.1a) \\ \tilde{\lambda}^k = \arg\min\big\{\dfrac{1}{2}(\lambda - \lambda^k)H_0(\lambda - \lambda^k) + (\tilde{s}_0^k)^T\lambda \mid \lambda \in \Lambda\big\}, & (6.1b) \\ \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \lambda^k \end{pmatrix} - \alpha \begin{pmatrix} x^k - \tilde{x}^k \\ \lambda^k - \tilde{\lambda}^k \end{pmatrix} \text{ with } \alpha \in (0,2), & (6.1c) \end{cases}$$

where $\tilde{s}_0^k = A(2\tilde{x}^k - x^k) - b$. Clearly, the scheme (6.1) includes the balanced ALM (2.2) as a special case with $\alpha = 1$. Numerically, the extra step (6.1c) has been shown to be able to accelerate the convergence of the classic PPA for various problems. We refer to, e.g., [3, 22, 24], for some empirical studies. Hence, it is motivated to consider the generalized scheme (6.1) to replace the balanced ALM (2.2).

To establish the convergence of (6.1), we just need to follow the roadmap in Section 3 and prove some similar theorems. For instance, the inequality (3.11) in Theorem 3.2 can be generalized as

$$\begin{aligned} &\alpha\big(\theta(x) - \theta(\tilde{x}^k) + (w - \tilde{w}^k)^T F(w)\big) \\ \geq\ & \frac{1}{2}\big(\|w - w^{k+1}\|_H^2 - \|w - w^k\|_H^2\big) + \frac{\alpha(2-\alpha)}{2}\|w^k - \tilde{w}^k\|_H^2, \ \ \forall w \in \Omega. \end{aligned}$$

Moreover, the inequality (3.14) in Theorem 3.3 can be generalized as

$$\|w^{k+1} - w^*\|_H^2 \leq \|w^k - w^*\|_H^2 - \alpha(2-\alpha)\|w^k - \tilde{w}^k\|_H^2, \ \ \forall w^* \in \Omega^*.$$

Then, based on these inequalities, analogous as the analysis in Section 3, convergence results for the generalized version of the balanced ALM (6.1) can be obtained trivially.

In addition, the extra step (6.1c) can be combined with the splitting versions of the balanced ALM (4.4) and (5.3) as well, and thus some generalized versions of the algorithms (4.4) and (5.3) can also be proposed. The details are omitted for succinctness.

7 Conclusions

In this paper, we reshape the classic augmented Lagrangian method (ALM) by balancing its subproblems. Convex programming problems with both linear equality and inequality constraints are considered. We propose a balanced ALM for the generic case, and various splitting versions for the separable cases. The balanced ALM and its splitting versions have the common feature that the subproblems are better balanced, and they are easier to be implemented for various applications. The balanced ALM advances the classic ALM by enlarging its applicable range, better balancing its subproblems, and improving its implementation. The balanced ALM and its splitting versions substantially enhance the rich literature of the classic ALM and its variants from a novel perspective, and open up the door to designing other application-tailored algorithms of the same kind for more specific/complicated problems.

References

[1] A. Beck, First-Order Methods in Optimization, MOS-SIAM Series on Optimization (2017).

[2] S. Becker, The Chen-Teboulle algorithm is the proximal point algorithm, manuscript, 2011, arXiv: 1908.03633[math.OC].

[3] D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods, Academic Press, New York (1982).

[4] E. J. Candes and B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math., **9** (2009), pp. 717-772.

[5] A. Chambolle, T. Pock, A first-order primal-dual algorithms for convex problem with applications to imaging, J. Math. Imaging Vison, **40** (2011), pp. 120-145.

[6] A. Chambolle and T. Pock, An introduction to continuous optimization for imaging, Acta Numer. **25** (2016), pp. 161–319.

[7] A. Chambolle and T Pock, On the ergodic convergence rates of a first-order primal-dual algorithm, Math. Program., **159** (2016), pp. 253-287.

[8] S. S. Chen, D. L. Donoho and M. A. Saunders, Atomic decomposition by basis pursuit, SIAM Rev., **43** (2001), pp. 129-159.

[9] R. W. Cottle , J. S. Pang and R. E. Stone, The Linear Complementarity Problem, SIAM, 2009.

[10] J. Eckstein, Some saddle-function splitting methods for convex programming, Optim. Meth. Soft., **4** (1994), pp.75-83.

[11] E. Esser, X. Zhang and T. F. Chan, A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, SIAM J. Imaging Sci., **3** (2010), pp. 1015-1046.

[12] M. Fortin and R. Glowinski, Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems. Amsterdam-New York, North-Holland Publ. Co. 1983.

[13] R. Glowinski and A. Marrocco, Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problémes non linéaires, RAIRO Anal. Numer. R2 (1975), pp. 41-76.

[14] R. Glowinski and P. Le Tallec, Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics. SIAM Studies in Applied Mathematics, Philadelphia, PA (1989)

[15] G. Golub and C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, The Fourth Edition, 2013.

[16] E. G. Gol'shtein and N. V. Tret'yakov, Modified Lagrangians in convex programming and their generalizations, Math. Program. Study, **10** (1979), pp. 86-97.

[17] B. S. He, A new method for a class of linear variational inequalities, Math. Program., **66** (1994), pp. 137–144.

[18] B. S. He, Solving a class of linear projection equations, Num. Math., **68** (1994), pp.71-80.

[19] B. S. He, L. Z. Liao, D. R. Han and H. Yang, A new inexact alternating directions method for monontone variational inequalities, Math. Program., **92** (2002), pp. 103-118.

[20] B. S. He, F. Ma and X. M. Yuan, Indefinite proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems, IMA J. Num. Anal., **75** (2020), pp. 361-388.

[21] B. S. He, F. Ma and X. M. Yuan, Optimally linearizing the alternating direction method of multipliers for convex programming, Comput. Optim. Appl., **75** (2020), pp. 361-388.

[22] B. S. He and X. M. Yuan, Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective, SIAM J. Imag. Science, **5** (2012), pp. 119-149.

[23] B. S. He and X. M. Yuan, On the $\mathcal{O}(1/n)$ convergence rate of Douglas-Rachford alternating direction method, SIAM J. Numer. Anal., **50** (2012), pp. 700-709.

[24] B. S. He and X. M. Yuan and W. X. Zhang, A customized proximal point algorithm for convex minimization with linear constraints, Comput. Optim. Appli., **56** (2013), pp. 559-572.

[25] M. R. Hestenes, Multiplier and gradient methods, J. Optim. Theory Appli, **4** (1969), pp. 303-320.

[26] Z. Lin, R. Liu and H. Li, Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning, Mach. Learn., **99(2)** (2015), pp. 287-325.

[27] B. Martinet, Regularisation, d'inéquations variationelles par approximations succesives, Rev. Francaise d'Inform. Recherche Oper., **4** (1970), pp. 154-159.

[28] J. Nocedal and S. J. Wright, Numerical Optimization, Second Edition, Springer, 2006.

[29] T. Pock and A. Chambolle, Diagonal preconditioning for first order primal-dual algorithms in convex optimization, in 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1762-1769.

[30] M. J. D. Powell, A method for nonlinear constraints in minimization problems, In Optimization edited by R. Fletcher, pp. 283-298, Academic Press, New York, 1969.

[31] B. Recht, M. Fazel and P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev., **52** (2010), pp. 471-501.

[32] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, Math. Oper. Res., **1** (1976), pp. 877-898.

[33] J. Stoer and R. Bulirsch, Introduction to Numerical Analysis, Springer, 2002.

[34] X. F. Wang and X. M. Yuan, The linearized alternating direction method of multipliers for Dantzig selector, SIAM J. Sci. Comput., **34** (2012), A2792–A2811.

[35] J. F. Yang and X. M. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, Math. Comp., **82** (2013), pp. 301-329.